AMERICAN JOURNAL OF PHARMTECH RESEARCH

Maximum Likelihood Estimators for the Generalized Yule Distribution

Amal T. Badawi
Department of Statistics, Faculty of Science, Jeddah University, Jeddah, Saudi Arabia.

ABSTRACT

Yule distribution is one of the more accurate distributions for fitting the heavy tailed data, although it is difficult to get a closed formula for the parameter estimator. In Spierdijk (2007), single maximum likelihood estimator (MLE) for the Yule (ρ) parameter was found numerically. In addition, another extension of the distribution was derived and it denoted as GYule (ρ,α), generalized Yule distribution. The moment estimators were got numerically for GYule but it was difficult to get the MLE's for the parameters. [Spierdijk (2007)]. In this study, single MLE's for GYule (ρ,α) parameters was found numerically and was applied for the same dataset used in Spierdijk (2007). GYule density function was also derived using the method of mixing distributions and then an explanation of variation was given by dividing the distribution variance into three components (randomness, liability and proneness).

Keywords: Yule distribution, generalized Yule distribution, extended Yule distribution, incomplete Beta function, mixed distributions, superstar data and the snowball effect.

*Corresponding Author Email: a_tolba_1@hotmail.com
Received 09 August 2018, Accepted 15 August 2018

Please cite this article as: Badawi AT., Maximum Likelihood Estimators for the Generalized Yule Distribution. American Journal of PharmTech Research 2018.

INTRODUCTION

Yule distribution is used to fit the heavy tailed data, which the frequencies of the small values are very large and it decreases strongly and exponentially as values increase, then the frequencies of the big values are very small, i.e. it contains a strong positive skewness. Examples of these data are: superstar data, phylogenetic trees, the frequency of words in novels and population in the geographical areas.

Some previous studies faced difficulty in estimating the Yule distribution parameter. The reason is the difficulty of dealing with the beta function which it included in the Yule density function. Then, it is difficult to get a closed form for the parameter estimator. This led to a discrepancy between some studies interested in applying Yule distribution; however they used the same dataset in some cases. The methods used in estimating Yule distribution parameter were not accurate. Some studies decided to find the Yule parameter estimation by approximating the Pareto distribution using the power law. Others decided to choose an arbitrary value for the parameter estimation ($\rho = 1$), based on that the best results were obtained was when putting ρ equal to one. Of course, they didn't have a realistic results but it was satisfactory for their research.

[Yule, G. (1925), Simon (1955), Adler, M. (1985), Chung, K.(1994,1998) , Cox, R. (2000), Giles, D. (2006)].

With regard to the ability of using Yule distribution to fit the data referred to, the controversy on this subject was discussed in [Spierdijk (2007)] where the Yule distribution was used by different way at this time. Pareto distribution can't be used to fit the data because of Pareto's inability for giving an accurate estimate of the lower quintiles for the distribution of the heavy tailed data as it gives exaggerated estimate of the snowball effect. Also, putting an arbitrary value for the parameter is completely unacceptable economically and mathematically.

Spierdijk (2007) used numerical methods of estimation by using some statistical programs that can handle a beta function in the estimation. The study concluded that the Yule distribution could fit the stardom variable, while it was unable to fit the superstardom variable. Thus, a new distribution of two parameters was obtained by replacing the complete beta function with the incomplete beta function in the Yule distribution density function. The study was applied on a real superstar data and concluded that the new distribution fitted the data better than the Yule distribution. The Problem in this study was the difficulty of getting the MLE's for the parameters, while the moment estimators were got numerically for GYule instead. The reason for this problem was the existence of some

numerical difficulties due to the emergence of the incomplete beta function in the logarithm of maximum likelihood function.

Some studies have been interested in obtaining other extensions for Yule distribution to make it more accuracy in fitting the heavy tail date. EYule(ρ, λ) presented in the Rodriguez (2011) study, is worth mentioning that they used the same data set that using in [Spierdijk (2007)]. The aim was to compare the results, but they also used the moment estimators for GYule instead of the MLE's.

In this study, single MLE's for GYule(ρ, α) parameters was found numerically, which was not exposed in any recent studies, and was applied for the same dataset used in [Spierdijk (2007), Rodríguez (2011)] and comparing the results with the other obtained in those studies.

GYule density function was also re-derived using the method of mixing distributions and then an explanation of variation was given by dividing the distribution variance into three components (randomness, liability and proneness).

GYule distribution

GYule(ρ, α) distribution is a generalization of Yule(ρ) distribution. Yule distribution presented in [Yule, G. (1925)] for the first time and re-presented in [Simon (1955)], so it was known as Yule-Simon distribution for a while. The density function of the random variable Y, which follows the Yule(ρ) distribution, is:

$$P[Y = y] = \rho B(y, \rho + 1) \quad ; \quad y = 1,2, \dots$$

Where ρ is the single parameter for the distribution, $(\rho > 0)$. And $B(a, b)$ is the Beta function with parameters *a* and *b*.

The study [Spierdijk (2007)] concerned what was referred to in the studies of [Yule (1924, sec. II)] and [Simon (1955, sec. I)] regarding the possibility of deriving another extension from the Yule distribution with two parameters. A new extension of the Yule distribution was obtained with two parameters and called a generalized Yule "GYule", where the density function of the random variable Y, which follows GYule(ρ, α) is:

$$P[Y = y] = \frac{\rho}{1 - \alpha^{\rho}} B_{1-\alpha}(y, \rho + 1) \; ; \; \alpha \in (0,1), \rho > 0 \; ; y = 1,2, \dots$$

Where:

- ρ and α are the distribution parameters, $(\rho > 0)$ and $(0 < \alpha < 1)$.
- $B_{\varepsilon}(a, b)$ is the incomplete Beta function,

$$B_{\varepsilon}(a, b) = \int_0^{\varepsilon} x^{a-1}(1 - x)^{b-1} \, . \, dx$$

- Yule distribution is a special case when $(\alpha = 0)$; i.e. $\text{Yule}(\rho) = \text{GYule}(\rho, 0)$. The first and second moments were found in the same study.

New definition for GYule distribution

In a study of [Rodríguez-Avi, (2007)], Yule distribution was found to belong to a family of Type I Gaussian hypergeometric distributions which it denoted as GHDI. Specifically, it is a special case of generalized Waring distribution which it denoted as $\text{UGWD}(a, k, \rho)$ when a=k=1, i.e. $[\text{Yule}(\rho) = \text{UGWD}(1,1,\rho)]$. The study of Rodríguez-Avi concluded that the density function of the Yule distribution can be derived by mixing the geometric distribution with the generalized beta distribution. This mixture is indicated by the following symbol:

$$\text{Geometric}(p) \bigwedge_{P} \text{BetaI}(\rho, 1)$$

Where $\text{BetaI}(\rho, 1)$ is the generalized beta distribution.

In this study another definition of the distribution of GYule was obtained by using the truncated generalized beta distributed instead of generalized beta distribution, i.e. using the mixture:

$$Geometric(P) \bigwedge_{P} BetaI(\rho, 1|\alpha)$$

Using this mixture, the same density function was obtained for GYule distribution through the following integration:

$$P[X = x] = \int_{\alpha}^{1} p(1-p)^{x} \frac{\rho p^{\rho-1}}{1-\alpha^{\rho}} \, .dp$$

Putting $\nu = (1-p)$ and make shift by put [Y =X +1], then:

$$P[Y = y] = \frac{\rho}{1-\alpha^{\rho}} B_{1-\alpha}(y, \rho+1) \quad ; \quad y = 1,2, \ldots$$

Although the same density function is obtained for the distribution of GYule, the derivation of the distribution in this way gives the distribution an important advantage, namely, the possibility of dividing the variation as described in the following subsection.

Dividing the variance of GYule distribution

The added advantage here is the possibility of obtaining distribution properties through the process of the mixture distributions. An explanation for the variance of GYule distribution was obtained by defining it as a result of mixing the geometric distribution, with the density function:

$$P[X|p] = p(1-p)^{x}, x = 0,1,2, \ldots \quad , \quad E_{X|P}[X|p] = \frac{1-p}{p}$$

With the truncated generalized beta distributed from below, with the density function:

$$f(p) = \frac{\rho p^{\rho-1}}{1-\alpha^{\rho}} \quad ; \quad \rho > 0 \, , 0 < \alpha < 1; \quad \alpha \leq p \leq 1$$

Then, the variance of GYule distribution was divided into three components according to the sources of variation as follow:

- Randomness (due to unknown factors*)*
- Liability (due to factors related to the observations themselves*)*
- Proneness (due to some factors not related the observations)

These components were shown in table (1). The total variance for GYule distribution is obtained by adding all of the three components mentioned above.

Table 1 dividing the variance of GYule distribution into three components

Source of	**Variance**
Randomness	$\frac{[1-\rho\alpha^{\rho-1}+(\rho-1)\alpha^{\rho}]}{(1-\alpha^{\rho})(\rho-1)}$
Liability	$\frac{2-\rho(\rho-1)\alpha^{\rho-2}+2\rho(\rho-2)\alpha^{\rho-1}-(\rho-2)(\rho-1)\alpha^{\rho}}{(1-\alpha^{\rho})(\rho-2)(\rho-1)}$
Proneness	$\text{Liability}-(\text{Randomness})^2$

MLE's for GYule distribution

In this study the maximum likelihood estimators for GYule distribution was obtained through maximization the following logarithm of the likelihood function:

$$\ln \mathrm{L} = n\ln\rho - n\ln(1-\alpha^{\rho}) + \sum_{i=1}^{\infty}\ln(B_{1-\alpha}(y,\rho+1))$$

Updated software can now calculate the incomplete beta function using numerical integrals, allowing using the maximum likelihood method of estimating the distribution parameters rather than using the moments method. In this study, R was used to obtain the MLE's numerically, and was applied to the same data set used in studies of [Spierdijk (2007), Rodríguez (2011)].

Application on the real data

The dataset used in this study is known in economics as superstar data, which is a description of superstar phenomenon. Superstar phenomenon occurs, for example, when there are relatively few individuals who gain huge incomes and thus control the labor market. Researchers differed in this field about the cause of this phenomenon. Some researchers attributed the reason for this phenomenon to the existence of differences talents for the individuals while others went to the view of study of [Adler (2006)], which concluded that this phenomenon can occur regardless of the talents for the individuals, and added a simple example of his point of view, saying that music has an important social aspect and many individuals are turning to the public opinion and then creating the

so-called snowball effect. (Snowball effect in art makes people record their voices to the most famous artists in the public, increasing their incomes more and more).

In study of [Chung, K. (1994)], the interested dataset was fitted using Yule distribution. These data are the Gold-Record Awards data, the source is the Recording Industry Association of America (RIAA) Inc. The study used a list of popular music stars during the period 1958-1989 and considered that the number of gold records was a measure of their artistic success. Then, [Spierdijk (2007) and Rodríguez (2011)] fitted the same dataset using the new extensions for Yule distribution which were introduced in their studies.

In this study the same dataset were also used to analyze and discuss the results with the previous studies clearly. The obtained results are shown in Table (2), which contains the results of data fitting using the GYule distribution (using the moments and maximum likelihood) and EYule (using maximum likelihood).

Table 2 the observed and predicted frequencies using GYule distribution (using both moments and maximum likelihood methods) and EYule (using maximum likelihood method) as well as the results of the χ^2 test and calculate the P-value's using Monte Carlo method (2000 times).

Gold	observed	GYule (MLE)	GYule (MM)	EYule(MLE)
1	668	658.04	632.23	654.87
2	244	246.31	250.5	248.23
3	119	131.99	138.43	133.38
4	78	82.49	87.98	83.25
5	55	56.19	60.41	56.53
6	40	40.44	43.57	40.53
7	24	30.23	32.51	30.18
8	32	23.23	24.86	23.11
9	24	18.23	19.37	18.08
10	14	14.55	15.32	14.38
11	16	11.77	12.26	11.6
12	13	9.63	9.91	9.47
13	11	7.95	8.07	7.8
14	5	6.61	6.63	6.48
15	4	5.53	5.47	5.43
16	4	4.66	4.54	4.57
17	2	3.95	3.78	3.87
18	7	3.36	3.17	3.29
19	2	2.86	2.66	2.81
20	3	2.45	2.24	2.41
21	1	2.11	1.89	2.08
22	3	1.82	1.6	1.79
23	1	1.57	1.36	1.55
24	1	1.36	1.16	1.35
25	0	1.18	0.99	1.17
26	0	1.03	0.84	1.02
27	0	0.89	0.72	0.9
28	0	0.78	0.62	0.78

29	1	0.68	0.53	0.69
30	0	0.60	0.46	0.6
31	0	0.52	0.39	0.53
32	0	0.46	0.34	0.47
33	0	0.40	0.29	0.41
34	1	0.36	0.25	0.37
35	0	0.31	0.22	0.32
36	1	0.28	0.19	0.29
37	1	0.24	0.16	0.26
38	0	0.22	0.14	0.23
39	0	0.19	0.12	0.2
40	0	0.17	0.11	0.18
41	0	0.15	0.09	0.16
42	0	0.13	0.08	0.14
43	0	0.12	0.07	0.13
44	0	0.10	0.06	0.11
45	1	0.09	0.05	0.1
46	1	0.08	0.05	0.09
> 46	0	0.67	0.32	0.8
Estimated parameters		$\hat{\rho} = 0.584$ (0.1) $\hat{\alpha} = 0.090$ (0.011)	$\hat{\rho} = 0.368$ (0.017) $\hat{\alpha} = 0.109$ (0.022)	$\hat{\rho} = 0.434$ (0.094) $\hat{\lambda} = 0.923$ (0.013)
Goodness of fit test		$\chi^2 = 53.132$ p = 0.247	$\chi^2 = 76.495$ p = 0.047	$\chi^2 = 51.619$ p = 0.275

It has been observed that for the positive values of the parameter ρ only the maximum likelihood of the GYule distribution can be obtained. The reason may regards to the constraints on the ρ parameter in the GYule distribution which it doesn't allow negative values for ρ unlike the case in EYule distribution, which allows ρ taking negative values.

In this study, an alternative test statistic for χ^2 was used by simulating the distribution of χ^2 using random sampling of the discrete distribution of the probabilities that were fitted. P-value was calculated using the Monte Carlo method used in the study of [Hope (1968)].

CONCLUSION

At 0.05 significant level, the Yule distribution was rejected to fit the data and both GYule and EYule were accepted to fit this dataset with very closed probabilities respectively [0.247, 0.275] but EYule distribution is the best. While a clear difference in the results was observed in GYule distribution using the two methods of estimation. In case of using moment method (MM), the result was to reject GYule distribution for fitting the data with probability [0.047] while it was accepted in the case of using of the maximum likelihood (ML), as it was summarized in table[2].

The variance for the two distributions for both extensions was also obtained by dividing into three components as shown in Table [3].

Table 3 the mean and the three component of the variance which fitted using MLE's for GYule and EYule distribution.

Data set	Distribution	Mean	Randomness	Liability	Proneness	Total Variance
Gold records	GYule(0.584,0.09)	3.20	2.20	10.52	5.68	18.40
	EYule(0.434,0.922)	3.20	2.20	10.69	5.84	18.73

The variance components of these data have been defined as follow: the variation due to external factors is the type of music presented, the conditions of marketing, advertising methods or other external factors. The variation due to the differences between the studied interested individuals is the talent for the artists.

It was observed that a finite variance of the Yule distribution couldn't be obtained for the data. The reason was that the parameter of Yule distribution was greater than one. This was also exactly what was stated in the study [Rodriguez (2011)]. It was noted that the higher variance component was due to external factors, in other words, the difference between the individuals was not random and it was not due to the difference in the talent of the artists themselves. This is consistent with the results concluded in the studies of [Spierdijk (2007) and Rodriguez (2011)], that the differences in success levels between artists didn't necessarily mean that there are differences between their talents, so the talent of the artist couldn't measure through the number of the awards they received. In general, it isn't possible to measure the artist's talent in terms of his fame or his income. This can be generalized in all fields (medical, economic, educational, etc.) and not only in the art field.

REFERENCES

1. Adler, M., (2006), "Stardom and talent". In: Ginsburgh, V.A., and Throsby, D. (Eds.), Handbook of the Economics of Art and Culture. Amsterdam: Elsevier, pp: 895-906.
2. Chung, K.H., Cox, R.A.K., (1994), "A stochastic model of superstardom: an application of the Yule distribution". Review of Economics and Statistics 76, p: 771-775.
3. Chung, K.H., Cox, R.A.K., (1998), "Consumer behavior and superstardom", Journal of Socio-Economics 27, pp: 263-270.
4. Cox, R.A.K., Kleiman, R.T., (2000), "A stochastic model of superstardom: evidence from Institutional Investor 's All-American Research Team", Review of Financial Economics 9, pp: 43-53.

5. Giles, D.E., (2006), "Superstardom in the US popular music industry revisited " Economics Letters 92, pp: 68-74.
6. Hope, A.C.A., (1968), "A simplified Monte Carlo significance test procedure". Journal of the Royal Statistical Society [B] 30, pp: 582-598.
7. Martínez-Rodríguez, A.M., Sáez-Castillo, A.J. and Conde-Sánchez, A., (2011), "Modelling using an extended Yule distribution", Computational Statistics and Data Analysis 55, pp: 863-873.
8. Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J. and Olmo-Jiménez, M.J., (2007), "A new generalization of the waring distribution", Computational Statistics and Data Analysis 51 (12), pp: 6138–6150.
9. Simon, H.A., (1955), "On a class of skew distribution functions", Biometrika 42, p: 425-440.
10. Spierdijk, L. and Voorneveld, M., (2007), "Superstars without talent? The Yule distribution controversy", Working Paper Series in Economics and Finance 658. Stockholm School of Economics.
11. Yule, G.U., (1925) "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis", F.R.S. Phil. Trans. Roy. Soc. Lond. B 213, pp: 21–87.